SELF ADAPTIVE METHODS FOR EVOLUTIONARY BASED SENSITIVE RULE HIDING

SYNOPSIS

Submitted for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

by

G.BHAVANI(Reg. No. 17PHEOF001)

Ph.D. Research scholar (Full Time)

Under the guidance of

Dr. S. SIVAKUMARI

Professor and Head

Department of Computer Science and Engineering

School of Engineering

Avinashilingam Institute for Home Science and Higher Education for Women

Coimbatore

October 2022

SELF ADAPTIVE METHODS FOR EVOLUTIONARY BASED SENSITIVE RULE HIDING

1. INTRODUCTION

Data Mining deals with extraction of unknown information from large amount of datasets. In recent days organizations are extremely dependent on data mining results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. This data includes sensitive data about individuals or organizations. While running data mining algorithm against such data, the algorithm not only extracts the knowledge but it also reveals the information which is considered to be private. The realthreat is that once information gets exposed to unauthorized party, it will be impractical to stop misuse. Privacy can be in danger when Data Mining techniques uses the identifiers which themselves are not very sensitive, but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information. Privacy is very important for trusted collaboration and interactions. Because of these privacy and data security concerns in data mining, the data owner hesitates while sharing data for data mining activities.

Privacy Preserving Data Mining (PPDM) is an area of research in data mining process to solve the above problem. Its ultimate goal is to allow one to extract relevant knowledge from large amount of data and provide accurate data mining result, while prevent sensitive information from disclosure or inference. PPDM techniques are categorized into data hiding, knowledge hiding and hybrid techniques. Data hiding can be done through data perturbation, cryptographic, anonymization and condensation techniques. Knowledge hiding can be done through association rule hiding and query auditing techniques. Combination of the two or more techniques results in hybrid techniques.

Association Rule Hiding (ARH) is considered to perform privacy preserving data mining. It refers to the process of modifying the original database in such a way that certain sensitive association rules disappear without seriously affecting the data and the non-sensitive rules. The primary goal is to protect access to sensitive information that can be obtained through non-sensitive data and inference rules. Here it is considered that it is not the data but the sensitive rules that create a breach to privacy. Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve

only those rules having support and confidence higher than the user specified minimum support and confidence threshold. So for hiding sensitive rules, the support or confidence of the sensitive association rules need to be decreased below the threshold value. There are different approaches developed to decrease the support of association rule for hiding association rule. Distortion is one such method done by the alteration of attribute value by a new value. Cuckoo Optimization Algorithm for Association Rule Hiding (COA for ARH) is a data distortion-based ARH technique which hides the sensitive association rules in the transaction database using the cuckoo search algorithm.

In this research work, Cuckoo Optimization Algorithm is improved using various methods. In the first phase, a fixed number of transactions were modified to hide the sensitive association rules to make COA suitable for different sized datasets. The multi-objective optimization problem raised by introducing fitness function is solved by introducing a Pareto-optimal solution to select the best results for association rule hiding. In the second phase, an item is chosen based on the correlation between sensitive and non-sensitive rules to lessen the side effects of non-sensitive rules. The selected item is eliminated and inserted in the transaction of each cuckoo based on objective parameters and sensitivity of transactions.In the third phase, by altering the value of the switching parameter and the total number of iterations the dataset is sanitized with a high convergence rate. Moreover, a balance between exploration and exploitation in QP-ICOA for ARH is achieved using crossover and mutation operators in the cuckoo search algorithm. In the fourth phase, MapReduce framework is introduced to reduce the computational time by parallelizing the BSE-QP-ICOA for the ARH process.

2. LITERATURE SURVEY

A hybrid perturbation technique used to perturb categorical and numeric attributes and the utility of the datasets is measured using maintenance of Rank of Features (CK) between the original and perturb datasets [1]. A novel approach based on the data distortion technique for ARH used to hide the sensitive items denoted in Representative Rules (RR) format [2]. The association rule was entirely supported by RR and was chosen to remove sensitive items to sanitize the transaction database. The algorithm for association rule hiding intersection lattice was developedto hide a particular set of association rules with the most negligible side effects and reduced complications [3]. Modified decrease support of right hand side item of rule clusters improves the speed of the mining process [4]. An improved

genetic algorithmused to wrap the sensitive association rules in a privacy-conserving manner [5]. Resultant and complete item sensitivities algorithms developed for wrapping sensitive association rules according to the consequent and total item sensitivities [6]. Border Rule based Distortion algorithm analysed and produce efficient results with minimum lost rules [7]. A rule hiding approachbased on evolutionary multi-objective optimization (EMO) proposed to preserve user data privacy [8]. COA optimally distorts the sensitive items to sanitize the original database through fixed number of transaction modifications [9]. A fuzzy logic approach was proposed to wrap sensitive association rules based on the sensitivity degree of each transaction in the dataset [10].

Whale optimization and Least Lion Optimization Algorithms (LLOA) presented to protect users' confidentiality through ARH [11]. The whale optimization algorithm mines the association rules for the given database and then using secret keys to sanitize the database using LLOA. A MAXARH algorithm [12] identifies the delicate rules and provides confidentiality. The Recursive Pattern Sanitization algorithm hides multiple sensitive itemsets irrespective of their size and support in a single parse of the closed patterns [13].Frequent Pattern growth algorithm was used to generate association rules and then modified genetic algorithm was used to hide sensitive association rules [14]. An efficient algorithm was developed to hide association rules using the new support formula [15]. The minimum support system, used in between the approaches for hiding association rules, served as the basis for the new support formula. The privacy-preserved minable data publication scheme (PMDP) to enable selective, sensitive association rules hiding while supporting the non-sensitive association rules mining [16]. A novel fitness function evaluated the utility of the data and also measured the quality of the solutions and privacy leakage.Dynamic bit vector data structure was integrated with the electromagnetic field optimization approach which produces some similar adjustments to increase the processing speed and reduce the demands on the solution storage space [17]. An optimized sanitization approach for minable data publication called SA-MDPwas a customized PSO algorithm that produced new particles by random mutation or learning from the best particle [18]. The proper fitness function was derived to guide the particles to run toward the optimal solution.Sanitization problem solved by developing heuristic path based on intuition to formulate it. Accuracy maximization problem was formulated as a nonlinear integer program and derive differentresults which supports to reduce the problem size [19].

3. PROBLEM DEFINITION

The convergence problems like premature convergence, population diversity; unbalanced exploration value exists in the previous algorithms. More number of iterations required to hide sensitive association rules makes COA unsuitable for large datasets.The hiding failure and lost rules on non-sensitive rules is still high while hiding the sensitive association rules. The COA has static variable limits which may leads to slow convergence rate to hide the sensitive rules. The COA in a particular nodule is inappropriate and it indicates with the problem of high computational cost when working with huge volume of data.

4.OBJECTIVES OF THE RESEARCH

- To handle the conflicts between multiple fitness functions and decrease the generation of lost rules and the hiding failure of sensitive rules through an improved optimization algorithm.
- To overcome the side-effects of non-sensitive rulesand increase the quality of preserving privacy with multiple right hand side and left hand side.
- To sanitize the dataset and speed up the convergence rate by balancing stochastic exploration.
- To optimize the computational time and memory consumption using Map Reduce framework.

5. RESEARCH METHODOLOGIES

The experiments were conducted with different sized datasets to show the suitability of ARH using the adult, bank marketing, and hardware store sales datasets. The adult dataset (Kaggle dataset) consists of 32,561 transactions and 14 items, with an average transaction length is 15. The bank marketing dataset (UCI Machine Learning Repository) consists of 4522 transactions and 17 items with an average transaction length of 17. The hardware store sales dataset is a real-time dataset collected about sales details from January 1, 2017, to January 1, 2018. It consists of 1,00,000 transactions and 1000 items with an average transaction length of 100.The existing and proposed association rule hiding algorithms implemented in Java JDK 1.6 and run on a Microsoft Windows 7 with an Intel processor running at 2.70 GHz and 4GB memory.

5.1 Improved Cuckoo Optimization Algorithm for Association Rule Hiding

Improved COAfor ARHalgorithm was used for differentsized datasets. The least amount of transactions for modifications is selected based on support and confidence threshold values. The preserving ability of COA for the ARH algorithm gets increased by introducing fitness function with multiple objectives. Because of using numerous purposes in the new fitness function of COA for ARH, there may be a chance for a multi-objective optimization problem. The multi-objective optimization problem is solved by introducing a Pareto-optimal solution to select the bestresults for association rule hiding. The performance of the Pareto-optimal solution is improved through Crowding Distance (CD) in the Pareto-optimal solution to choose the optimal set of keys. Improved COA for ARH-CD (ICOA for ARH-CD) suits assorted datasets and efficiently hides the sensitive rules with lesser side effects.

5.2 Quality Preserving Improved Cuckoo Optimization Algorithm for Sensitive Rule Hiding

Quality Preserving ICOA for ARH (QP-ICOA for ARH) hides the sensitive rules with multiple right and left-hand sides. In QP-ICOA for ARH, to lessen the side effects of non-sensitive rules, an item is chosen based on the correlation between sensitive and non-sensitive rules. The selected item is eliminated and inserted in the transaction of each cuckoo based on objective parameters and sensitivity of transactions. This procedure is continued until a non-dominated best solution for association rule hiding is attained. This result in hiding the sensitive association rules by eliminating and reinserting items in the transactions through each cuckoo. Removing and inserting items in the transactions reduces the side-effect of non-sensitive rules.

5.3 Balancing Stochastic Exploration in Quality Preserving Improved Cuckoo Optimization Algorithm for Sensitive Rule Hiding

In Balancing Stochastic Exploration in QP-ICOA for ARH, variable limits are adjusted dynamically according to the fitness value of each cuckoo in the population. It also helps to sanitize the database with a high convergence rate. The efficiency of QP-ICOA for ARH is improved by dynamically changing the value of the switching parameter and the total number of iterations. Moreover, a balance between exploration and exploitation in QP-ICOA for ARH is achieved using crossover and mutation operators in the cuckoo search algorithm.

At every iteration, the population size is modified by the linear population reduction method that helps to increase the convergence rate and escape from the local optima problem.

5.4 MapReduce Framework for Association Rule Hiding

The BSE-QP-ICOA for ARH is a costlier process when applied to large datasets. So by introducing MapReduce framework the computational time can be reduced. The MapReduce framework implements the BSE-QP-ICOA for ARH in a distributed environment. Initially, the transaction database is fragmented into several accessible portions and given as input to the map function. In the map function, the least number of transactions for alterations is carefully chosen, and the sensitive items in those transactions are altered based on each cuckoo's fitness function. As a final point, the map function reciprocated the fitness function of every cuckoo. It was assigned as input towards the reducer function, which decided on the most proper fitness function. Based on the adequate fitness function, the cuckoo in the map function transfers the location and cleans up the transaction database. The entire method is Parallelized BSE-QP-ICOA for ARH (PBSE-QP-ICOA for ARH). This process results in reduced computational time by parallelizing the BSE-QP-ICOA for the ARH process using MapReduce.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

The association rule hiding techniques are compared from existing COA for ARH with ICOA -CD, QP-ICOA, BSE-QP-ICOA and PBSE-QP-ICOAand analyzed in terms of hiding failure, lost rule, memory consumption and computational time.

6.1 Hiding Failure

Hiding Failure (HF) denotes the number of sensitive rules which the sanitization algorithm could not hide and are still mined from the sanitized data.

$$Hiding\ Failure = \frac{|R_s(D')|}{|R_s(D)|} \qquad (6.1)$$

Where, $R_s(D')$- Number of sensitive rules discovered in D'

$R_s(D)$- Number of sensitive rules discovered in D.

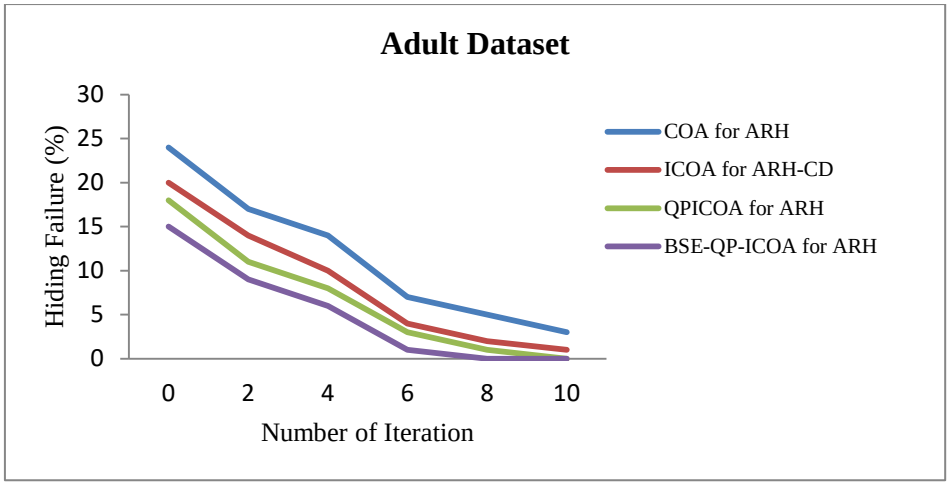

Figure 6.1 Comparison of hiding failure on adult dataset

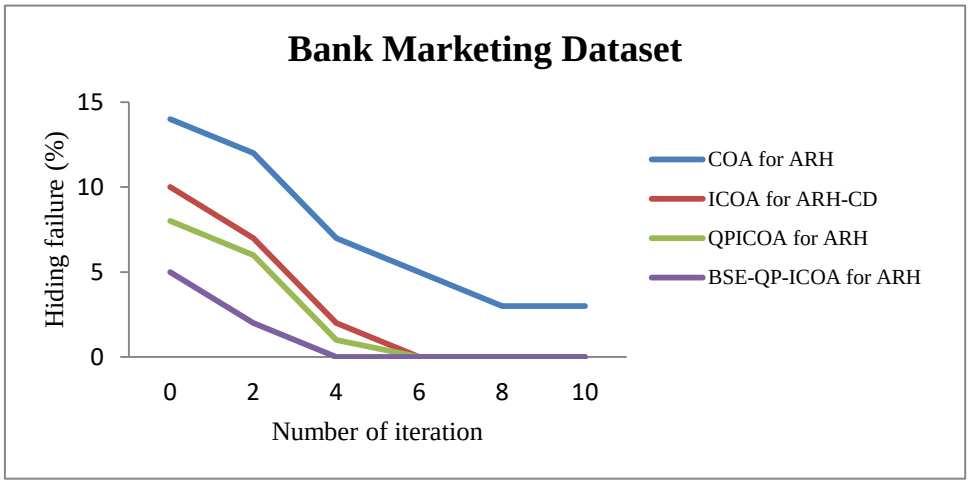

Figure 6.2 Comparison of hiding failure bank marketing dataset

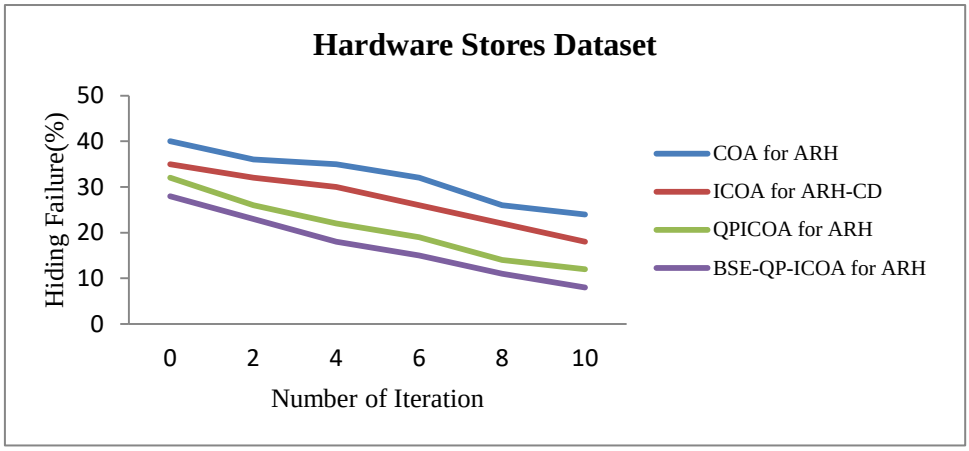

Figure 6.3 Comparison of hiding failure on hardware store dataset

From figures 6.1, 6.2 and 6.3 it is clear that the BSE-QP-ICOA for ARH outperforms other algorithms and results in minimum hiding failure than the COA for ARH, ICOA for ARH and QP-ICOA for ARH algorithms for all the three datasets.

6.2 Lost Rule

Lost Rule (LR) denotes the number of non-sensitive rules lost because of the act of associationrule hiding methods.

$$Lost\ Rule\ = \frac{|\sim R_S(D)| - |\sim R_S(D')|}{|\sim R_S(D)|} \quad (6.2)$$

Where, $|\sim R_S(D)|$- Number of non-sensitive rules explored in D

$|\sim R_S(D')|$- Number of non-sensitive rules explored inD'.

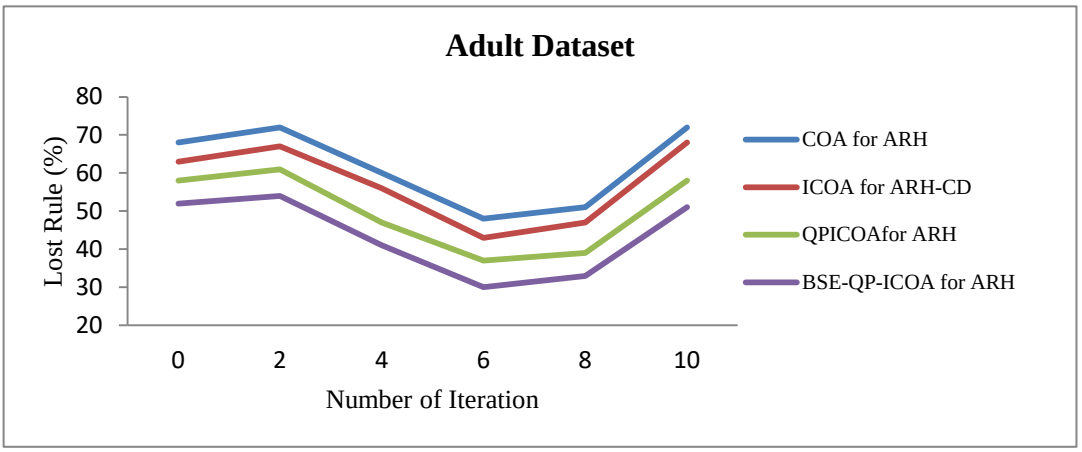

Figure 6.4 Comparison of Lost Rules on adult dataset

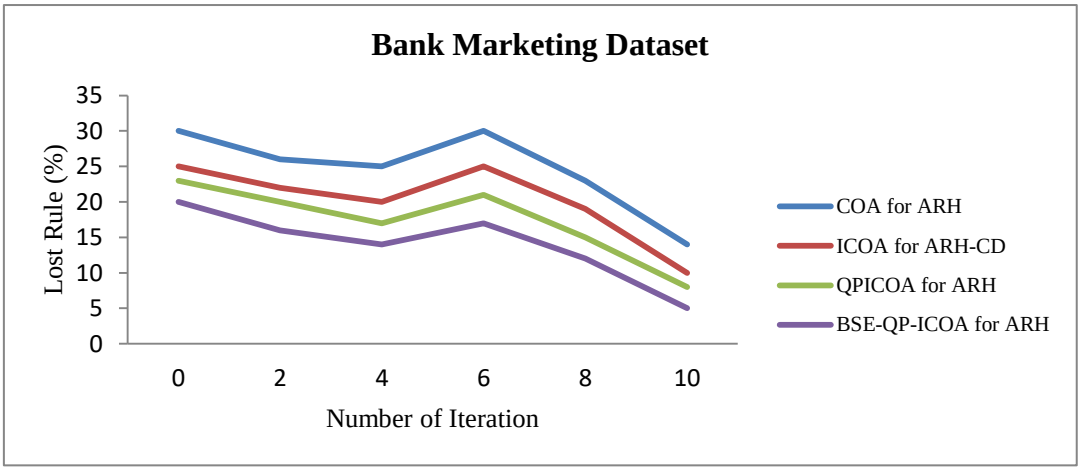

Figure 6.5 Comparison of Lost Rules on bank marketing dataset

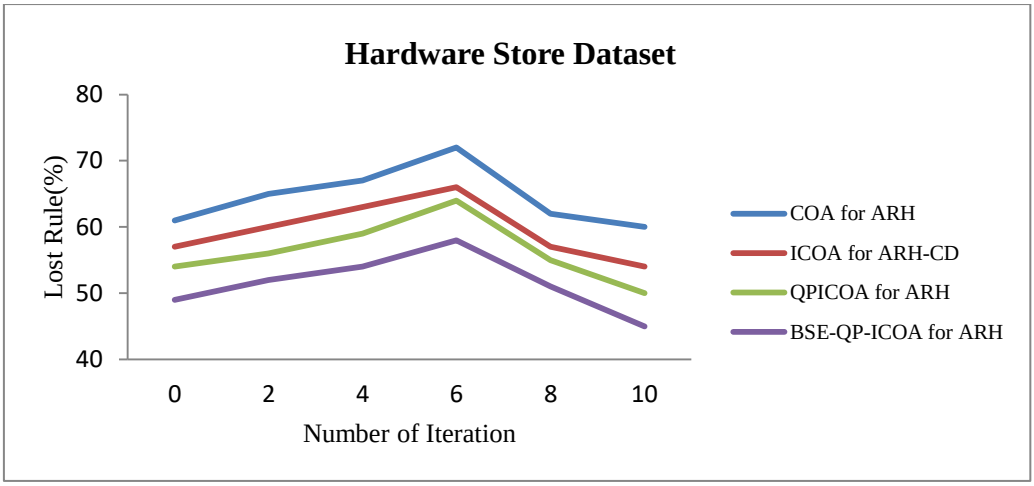

Figure 6.6 Comparison of Lost Rules on bank marketing dataset

From figures 6.4, 6.5 and 6.6 it is clear that the BSE-QP-ICOA for ARH outperforms other algorithms and reduces lost rules than the COA for ARH, ICOA for ARH and QP-ICOA for ARH algorithms for all the three datasets.

6.3 Memory Consumption

It denotes the amount of memory consumed by the existing and proposed method for sanitizing the transaction database.

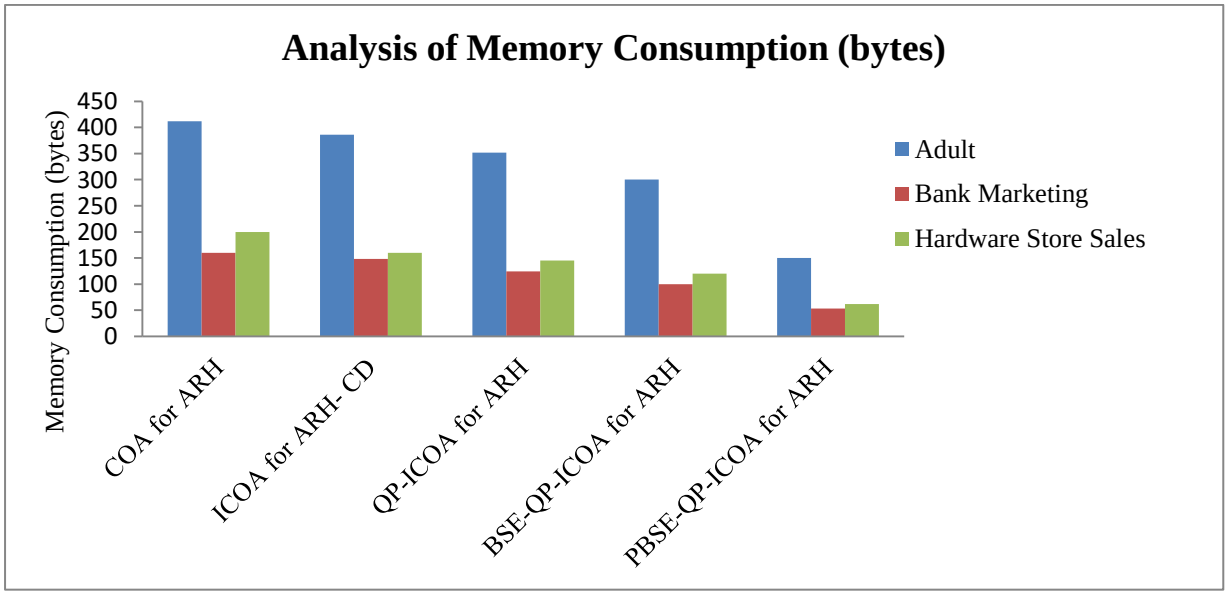

Figure 6.7 Comparison of memory consumption for Different ARH algorithms

From figure 6.7, it is clear that the PBSE-QP-ICOA for ARH outperforms other algorithms and results in reduced memory consumptionthan the COA for ARH, ICOA for ARH, QP-ICOA for ARH and BSE-QP-ICOA for ARH algorithms for all the three datasets.

6.4 Computational Time

It represents the time taken by association rule-hiding methods to sanitize the transaction database.

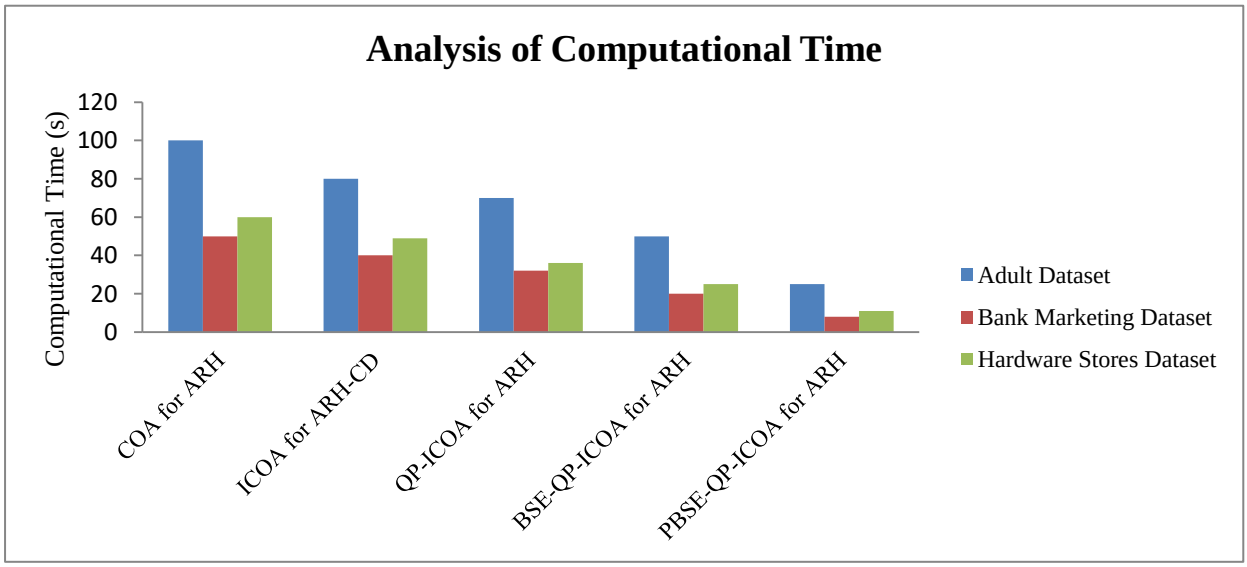

Figure 6.8 Analysis of Computational time for different ARH algorithms

Figure 6.8 show that the PBSE-QP-ICOA algorithm has better memory consumption and computational time for three different datasets. The performance analysis of ARH algorithms concludes that the BSE-QP-ICOA has a better hiding failure and lost rule. PBSE-QP-ICOA has better memory consumption and computational time than COA for ARH, ICOA for ARH and QP-ICOA for ARH algorithms.

7. CONCLUSION

This research focuses on hiding the sensitive association rules in the transaction database with fewer side effects. After initializing a minimum set of transactions for ARH, the multi-objective optimization problem in COA for ARH was solved by proposing ICOA for ARH-CD. Also, the side effects of non-sensitive rules while sanitizing the transaction database are reduced by using QP-ICOA for ARH. The performance of QP-ICOA for ARH is further enhanced by adjusting the variable limits dynamically using BSE-QP-ICOA for ARH. It also maintains a trade-offbetween exploration and exploitation using crossover and mutation operators in the cuckoo search algorithm. Finally, BSE-QP-ICOA for ARH is implemented in a distributed environment using the MapReduce framework to decrease the computational time and storage for ARH. The experimental results proved that the PBSE-QP-

ICOA for ARH has superior computational time and memory consumption performance than BSE-QP-ICOA for ARH. Also, it inherits the advantage of BSE-QP-ICOA for ARH, such as reduced hiding failure and fewer lost rules than other algorithms for ARH.

8. ORGANIZATION OF THE THESIS

Chapter 1 describes the detailed information about the PPDM, ARH, ARH techniques, their merits and demerits, motivation and contribution of this research work. **Chapter 2** studies and analyses the literature surveys which are related to this research work. **Chapter 3** explains the first objective of our research work such as ICOA for ARH-CD for ARH. **Chapter 4** describes the second objective of the research such as QP-ICOA for ARH to reduce the side-effect on non-sensitive rules. **Chapter 5** presents the third objective of the research such as BSE-QP-ICOA to reduce the speed up the convergence rate of QP-ICOA for ARH. **Chapter 6** describes the fourth objective of the research such as PBSE-QP-ICOA to reduce the computational time for ARH. **Chapter 7** explains the complete performance analysis of the projected methods. **Chapter 8** concludes the research work and provides the possibility of upcoming development.

REFERENCES

[1] Priyadarsini, P., Valarmathi, M. L., &Sivakumari, S. (2012). Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining. *International Journal of Computer Applications, 58(2).34-41*.

[2] Gulwani, P. (2012). A Novel Approach for Association Rule Hiding. *International Journal of Advance Innovations, Thoughts & ideas, 1(3), 1-9.*

[3] Quoc Le, H., Arch-int, S., & Arch-int, N. (2013). Association rule hiding based on intersection lattice. *Mathematical Problems in Engineering, 2013. 1-11*.

[4] Khan, A., Qureshi, M. S., &Hussain, A. (2014). Improved genetic algorithm approach for sensitive association rules hiding. *World Applied Sciences Journal, 31(12), 2087-2092.*

[5] Ponde, P. R., & Jagade, S. M. (2014). Privacy-Preserving by Hiding Association Rule Mining from Transaction Database. *IOSR Journal of Computer Engineering (IOSR-JCE), 16(5), 25-31*.

[6] Shahsavari, A., &Hosseinzadeh, S. (2014). CISA and FISA: efficient algorithms for hiding association rules based on consequent and full item sensitivities. In *7'th International Symposium on Telecommunications (IST'2014)*(pp. 977-982).

[7] Cheng, P., Lee, I., Pan, J. S., Lin, C. W., & Roddick, J. F. (2015). Hide association rules with fewer side effects.*IEICE TRANSACTIONS on Information and Systems, 98(10), 1788-1798.*

[8] Cheng, P., Lee, I., Lin, C. W., & Pan, J. S. (2016). Association rule hiding based on evolutionary multi-objective optimization. *Intelligent Data Analysis, 20(3), 495-514.*

[9] Afshari, M. H., Dehkordi, M. N., &Akbari, M. (2016). Association rule hiding using cuckoo optimization algorithm.*Expert Systems with Applications, 64, 340-351.*

[10] Afzali, G. A., &Mohammadi, S. (2017). Privacy-preserving big data mining: association rule hiding using fuzzy logic approach. *IET Information Security, 12(1), 15-24.*

[11] Menaga, D., &Revathi, S. (2018). Least lion optimization algorithm (LLOA) based secret key generation for privacy-preserving association rule hiding. *IET Information Security, 12(4), 332-340.*

[12] Murthy, T. S., Gopalan, N. P., &Venkateswarlu, Y. (2018). An Efficient Method for Hiding Association Rules with Additional Parameter Metrics. *International Journal of Pure and Applied Mathematics, 118*(7), 285-290.

[13] Surendra, H., Mohan, H.S., (2019). Hiding sensitive itemsets without side effects. *Applied Intelligence 49(4), 1213-1227.*

[14] Patel, J., & Shah, P. (2019). Hiding Sensitive Association Rules Using Modified Genetic Algorithm: Subtitle as needed (paper subtitle). In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 30-34). IEEE.

[15] Geethamary, A., Reddy, A.M.S., Agarwal, K. &Sawa, S. (2020). Optimized support balance model for sensitive rule hiding. *International Journal of Scientific and Technology Research, 9(4):2476–2480.*

[16] Yang, F., Lei, X., Le, J., Mu, N., & Liao, X. (2021). Minable data publication based on sensitive association rule hiding. *IEEE Transactions on Emerging Topics in Computational Intelligence. 14(8), 1-11.*

[17] Le, B., Le, D. P., & Tran, M. T. (2021). Hiding sensitive association rules using the optimal electromagnetic optimization method and a dynamic bit vector data structure. *Expert Systems with Applications, 176, 114879.*

[18] Yang, F., & Liao, X. (2022). An Optimized Sanitization Approach for Minable Data Publication. *Big Data Mining and Analytics, 5(3), 257-269.*

[19] Menon, S., Ghoshal, A., &Sarkar, S. (2022). Modifying Transactional Databases to Hide Sensitive Association Rules. *Information Systems Research, 33(1), 152-178.*

LIST OF PUBLICATIONS

1. Bhavani, G., &Sivakumari, S. (2019). Improved Cuckoo Optimization Algorithm for Association Rule Hiding.*International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol. 8, Issue - 10, 337-342. (SCOPUS)*
2. Bhavani, G., &Sivakumari, S. (2019). Quality Preserving ICOA for ARH. *International Journal on Emerging Technologies (IJET), ISSN (Print): 0975-8364, ISSN (Online):2249-3255, Vol.10, Issue – 4, 472-477. (SCOPUS)*

LIST OF CONFERNCES ATTENDED

1. "A survey on preserving data in association rule hiding" on International Conference on Intelligent Computing, Nano Science and Renewable Energy sources at Selvam College of technology, Namakkal on 08.03.2018 and 09.03.2018.
2. "Quality Preserving Improved Cuckoo Optimization Algorithm for Association Rule Hiding" on AICTE sponsored National Conference on Recent Advances in Computing (NCRAC- 2019) at Sri Krishna College of Engineering and Technology, Coimbatore on 19.09.2019 and 20.09.2019.